Grundkonzepte der explorativen Faktorenanalyse

Thome, Helmut

Historical Social Research, Vol. 16 — 1991 — No. 3, 3-39

Grundkonzepte der explorativen Faktorenanalyse[1]

*Helmut Thome**

Abstract: Most »Introductions« into factor analysis fail to give a detailed account of how the analytic task of constructing unobserved »factors« from observed variables translates into the mathematical problem of finding the Eigenvalues and Eigenvectors of the correlation matrix. Presupposing only elementary knowledge of matrix operations and differential calculus, this article attempts to fill this gap thereby enhancing the understanding of various facets of factor analysis. Two basic techniques, principal components and principal factor analysis, are applied to socio-economic indicators of Schleswig-Holstein communities in 1840.

1. Vorbemerkung

Die Faktorenanalyse ist ein mathematisch ziemlich aufwendiges Verfahren, das ohne elementare Kenntnis der Matrizen- und der Differentialrechnung kaum zu verstehen ist. Aber selbst wenn man sich als Soziologe oder Historiker gewisse Kenntnisse auf diesen Gebieten angeeignet hat, bleibt man bei der Lektüre einschlägiger Texte zur Faktorenanalyse selten vor schlimmen Frustrationen bewahrt. Entweder fehlen gänzlich diejenigen Argumentationsschritte, die eine Anwendung der Matrizen- und Differentialrechnung erfordern (was notwendigerweise zu Verständnislücken führen muß), oder der Leser wird mit Mathematik so stark überfordert, daß er den roten Faden verliert. So bleibt z. B. sehr häufig unklar, wie die analytische Aufgabe der »Faktorenextraktion« mit der mathematischen Theorie der Eigenwerte und Eigenvektoren verbunden ist. Wer sich nur für »Anwendungen« interessiert, mag darob nicht bekümmert sein. Sicherlich behindert eine solche Verständnislücke aber die kompetente Ent-

* Address all communications to Helmut Thome, Zentrum für Historische Sozialforschung, Zentralarchiv für Empirische Sozialforschung, Universität zu Köln, Bachemerstr. 40, D-5000 Köln 41.

Scheidung zwischen verschiedenen Modell- und Verfahrensalternativen, und sie erschwert auch die Interpretation der »Ergebnisse«, wie sie in Computerausdrucken präsentiert werden.

Das vorliegende Papier soll helfen, derartige Verständnisprobleme zu überwinden - unter der Voraussetzung, daß der Leser mit der Regressionsanalyse vertraut ist und über besagte Elementarkenntnisse der Matrizen- und Differentialrechnung verfügt (das Eigenwertproblem wird hier aber noch einmal skizziert). Es werden nur Grundzüge (keine verfahrenstechnischen Details) der explorativen (deskriptiven) Faktorenanalyse erläutert. Der allgemeine modelltheoretische Ansatz, in dessen Rahmen hypothesentestende Verfahren und statistische Schätzmethoden zu begründen sind, wird hier nicht dargestellt - siehe dazu z. B. die Einführung von Bacher 1990, Kap. 5. Der hier vorgelegte Artikel soll den Zugang zu Baders Arbeit erleichtern.

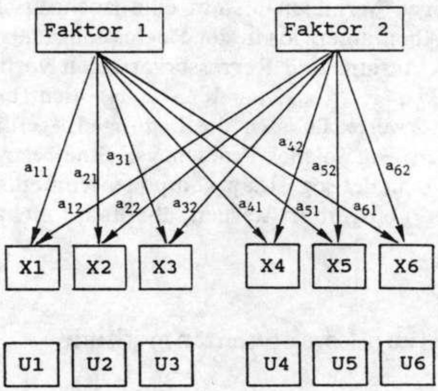
2. Das Ausgangsproblem

Machen wir uns zunächst die typische Aufgabenstellung klar, die der Forscher mit Hilfe der Faktorenanalyse bearbeiten möchte. Ausgangspunkt ist eine Menge von Meßergebnissen für einen Satz von K metrischen[2] Variablen X_k (k=l,2,...,K) bei N Untersuchungseinheiten (UE_i, i=l,2,...,N). Beispielsweise hat der Soziologe die Meinungen von N Personen zu K politischen und alltäglichen Streitfragen registriert; oder ein Sozialhistoriker hat für N Staaten oder Regionen K Indikatoren zum gesellschaftlichen Entwicklungsstand Mitte des 19. Jahrhunderts gesammelt. Die Menge dieser Informationen soll nun reduziert werden, ohne den Informationsgehalt allzusehr zu beeinträchtigen. Der Soziologe möchte die Vielzahl der Meinungsäußerungen in wenigen Einstellungsdimensionen, der Sozialhistoriker die verschiedenen gesellschaftlichen Entwicklungsindikatoren zu einigen Dimensionen der »Modernisierung« zusammenfassen. Aus den Beobachtungsdaten sollen also $Q \leqq K$ neue Skalen (»latente Dimensionen« oder »Faktoren«) F_q (q=l,2,...,Q) so konstruiert werden, daß mit ihrer Hilfe die Werte x_{ik} der Ausgangsvariablen X_k bis auf eine vernachlässigbare Restgröße reproduziert werden können. Wenn die Faktoren ermittelt sind, soll für eine beliebige UE_i folgendes Gleichungssystem gelten:

$$
\begin{aligned}
(1)\quad x_{1i} &= c_1 + a_{11}f_{1i} + a_{12}f_{2i} + \ldots + a_{1q}f_{Qi} + u_{1i} \\
x_{2i} &= c_2 + a_{21}f_{1i} + a_{22}f_{2i} + \ldots + a_{2q}f_{Qi} + u_{21} \\
&\ldots \\
X_{ki} &= c_k + a_{K1}f_{1i} + a_{K2}f_{2i} + \ldots + a_{kq}f_{Q}i + u_{K1}
\end{aligned}
$$

Mit dem Gleichungssystem (1) bewegen wir uns im Rahmen linearer Modelle. Jeder UE ist auf jedem Faktor ein bestimmter Wert zugewiesen. Jeder Faktor wird variablen spezifisch mit einer bestimmten Größe a_{kq} gewichtet. Diese Gewichtskoeffizienten bezeichnet man als Faktorladungen (»factor loadings«). Sie sind zunächst genauso unbekannt wie die Faktoren. Die Faktoren F_q sind, so wird im Basismodell angenommen, allen Ausgangsvariablen gemeinsam (»gemeinsame Faktoren«, »common factors«), aber die Variablen unterscheiden sich in den jeweiligen Faktorladungen. In der Regel strebt man eine Lösung an, in der unterschiedliche Gruppen von Variablen jeweils nur auf einem (jede Gruppe auf einem anderen) Faktor hoch »laden«. Die Ausdrücke u_{ki} repräsentieren Restgrößen in X_k, die nicht mit Hilfe der gemeinsamen Faktoren reproduziert (»erklärt«) werden können. In der Regel geht man davon aus, daß diese Restgrößen (»Einzelrestfaktoren«), die auch Meßfehler enthalten können, weder mit den gemeinsamen Faktoren noch untereinander korrelieren. Sie gelten jeweils als spezifische Einflußgrößen (»unique factors«) für eine bestimmte Variable X_k. Das allgemeine Modell läßt sich für den Fall zweier Faktoren mit Hilfe eines Pfaddiagramms wie folgt veranschaulichen (s. Abb. 1).

Abb. 1: Pfaddiagramm zum Faktorenmodell



Das Verhältnis der Faktoren (wenn man sie erst einmal »gefunden« hat) zu den einzelnen Variablen kann theoretisch auf unterschiedliche Weise interpretiert werden. Der Umfrageforscher z. B. mag die Faktoren als »Ein-

Stellungen« oder »Werte« interpretieren, die das beobachtete Antwortverhalten bis auf den unaufgeklärten Rest »verursachen«. Der Sozialhistoriker mag nach allgemeinen Dimensionen gesellschaftlicher »Modernisierung« suchen, die in Indikatoren wie »Anteil der industriell Beschäftigten« oder »Anteil der Bevölkerung mit mehr als vier Jahren Schulbesuch« ihren »Ausdruck« finden, ohne die begrifflich abstrakteren Modernisierungsdimensionen substanztheoretisch als »Ursachen« der beobachteten Merkmale deuten zu wollen.[3] In noch anderen Situationen mag es um die Informationsreduktion an sich gehen, ohne daß die neuen Skalen theoretisch gedeutet werden.[4] Mit dem Modell (1) können sich also durchaus unterschiedliche Absichten verbinden, die sich teilweise in unterschiedlichen Zusatzannahmen (Modellrestriktionen) niederschlagen.

In welchem Maße die in den beobachteten Variablen enthaltene Information in den Faktoren bewahrt bleibt, läßt sich an dem Anteil der Varianzen der X-Variablen ermessen, den die Faktoren in einem technischen Sinne zu »erklären« vermögen. Wie gut sich die einzelnen latenten Dimensionen (u. U. gibt es auch nur eine) analytisch voneinander trennen lassen, hängt von dem Muster der Faktorladungen ab (wird weiter unten erläutert). Wenn keine Hypothesen über die Zahl der Faktoren und die Zuordnung der einzelnen Variablen zu den latenten Dimensionen vorliegen, versucht man mit Hilfe einer »explorativen« Analyse eine Faktorstruktur anhand der beobachteten Daten zu »entdecken«.

Das Gleichungssystem (1) erinnert an die Regressionsanalyse, bei der man mit Hilfe eines Satzes von optimal gewichteten Regressorvariablen die Werte einer abhängigen Variablen bis auf eine stochastische Fehlergröße voraussagen will. Aber anders als in der Regressionsanalyse, bei der beobachtete Werte für Kriterium und Regressorvariablen vorliegen, sind hier die Faktorwerte f_q (q=1, 2, ..., Q; i=1, 2, ..., N) nicht bekannt. Es müssen sowohl die Faktorwerte als auch die Ladungen »gefunden«, genauer: konstruiert werden. Ein solches Problem ist ohne restriktive Modellannahmen, wie sie z. B. in der sog. Hauptkomponentenmethode (»principal component analysis«) eingeführt werden, überhaupt nicht lösbar.

3. Die Hauptkomponentenmethode

Wir wollen zunächst alle beobachteten Variablen z-standardisieren (was keine Einschränkung der Allgemeinheit bedeutet):

(2) $z_{ki} = (x_{ki} - \bar{x}_k)/s_{x(k)}$,

wobei $\bar{x}_k$ das arithmetische Mittel und $s_{x(k)}$ die Standardabweichung der Variablen X_k darstellen.

Aus Gleichung (1) wird also

$$(3) \quad z_{ki} = \sum_{q=1}^{Q} (a_{kq} f_{qi}) + u_{ki} \qquad \text{für } k=1,2,\ldots,K;\ i=1, 2, \ldots, N$$

Die Z-Standardisierung hat zunächst den Vorteil, daß man die Summe der Varianzen in den K X-Variablen, die durch die Q Faktoren möglichst weitgehend »erklärt« werden soll, sofort angeben kann: sie ist gleich der Anzahl K, denn nach der Standardisierung weist jede Variable eine Varianz von 1 auf. Wenn wir außerdem das Konstruktionsverfahren so gestalten können, daß auch die Faktorwerte f_{qi} z-standardisiert und die Faktoren nicht korreliert sind, ergibt sich für die Faktorladungen eine anschauliche Interpretation: Die Gewichte a_{kq} sind standardisierte Regressionskoeffizienten unkorrelierter Regressoren; das Quadrat $(a_{kq})^2$ ist somit derjenige Varianzanteil der Variablen Z_k, der durch den Faktor F_q »erklärt« werden kann. Entsprechend gibt die Summe der Ladungsquadrate

$$\sum_{q=1}^{Q} (a_{kq}) = (h_k)^2 \leqq 1\ , \ k=1,2,\ldots,K$$

an, wie groß der durch die gemeinsamen Faktoren erklärte Varianzanteil der Variablen Z_k ist. Diese Größe bezeichnet man als »Kommunalität«. Die Summe

$$\sum_{k=1}^{K} (a_{kq})^2 = V_q\ , \ q=1,2,\ldots,Q$$

gibt dagegen an, wieviel Varianz ein einzelner Faktor F_q insgesamt, über alle Variablen X_k betrachtet, erklärt. Wie wir später sehen werden, wird diese Größe durch den sog. Eigenwert des jeweiligen Faktors repräsentiert.

Wir wollen nun den Lösungsweg skizzieren, das heißt vor allem, die Modellannahmen erläutern, die es ermöglichen, Faktoren und Faktorladungen zu bestimmen. Der einfachste Ansatz, die sog. Hauptkomponentenmethode, beruht im wesentlichen auf drei Vorgaben: Als erstes gehen wir davon aus, daß die Faktoren, wie schon erwähnt, nicht miteinander korrelieren.[5] Zweitens setzen wir voraus, daß es keine spezifischen Einzelrestfaktoren, sondern nur gemeinsame Faktoren gibt, die die gesamte Varianz der Z-Variablen reproduzieren. Formal bedeutet dies, daß wir für die Faktoren zunächst die gleiche Anzahl vorsehen wie für die beobachteten Variablen: Q = K. Diese vielleicht merkwürdig erscheinende Annahme (merkwürdig, weil wir ja an einer Datenreduktion interessiert sind) wird dadurch akzeptabel, daß wir, drittens, die Faktoren schrittweise wie folgt bestimmen: Ein erster Faktor wird so abgeleitet, daß er eine maximale Varianzsumme in den erhobenen Variablen X_k (k=1,2,...,K) erklärt; der zweite Faktor soll dann einen maximalen Anteil der Restvarianz reproduzieren usw. Man kann später überlegen, ob einige der zuletzt extrahierten Faktoren so wenig Varianz erklären, daß sie eliminiert werden können. Aber in der ersten Stufe besteht die Hauptkomponentenanalyse darin, daß ein Satz von Variablen, die miteinander korrelieren, in einen Satz von Variablen (»Faktoren«) transformiert wird, die untereinander nicht korrelieren. Wären alle Variablen X_k (k=l,2,...,K) miteinander unkorreliert, wären die Faktoren identisch mit den Variablen. Im entgegengesetzten Extremfall, wenn alle X-Variablen perfekt miteinander korrelierten, wäre die gesamte Varianz durch einen einzigen Faktor reproduzierbar, der mit einer beliebigen X-Variablen identisch wäre.

Mit Hilfe dieser Annahmen werden, ausgehend von Gleichung (1) bzw. (3), folgende Ableitungen möglich: Da wegen Q=K die gesamte Varianz der X- bzw. Z-Variablen durch die Faktoren erklärt wird, die Restgrößen also entfallen, wird aus Gleichung (3) die Gleichung

$$(4) \qquad z_{ki} = \sum_{q=1} (a_{kq} f_{qi}) \quad \text{für } k=1, 2, ..., K;\ i=1, 2, ..., N;$$

oder in Matrixschreibweise, wenn wir die Gleichungen für jede UE untereinander schreiben:

$$(5) \qquad \begin{pmatrix} z_{11}, & z_{12}, \dots, & z_{1K} \\ & \dots & \\ z_{i1}, & z_{i2}, \dots, & z_{iK} \\ & \dots & \\ z_{N1}, & z_{N2}, \dots, & z_{NK} \end{pmatrix} = \begin{pmatrix} f_{11}, & f_{12}, \dots, & f_{1Q} \\ & \dots & \\ f_{i1}, & f_{i2}, \dots, & f_{iQ} \\ & \dots & \\ f_{N1}, & f_{N2}, \dots, & f_{NQ} \end{pmatrix} \cdot \begin{pmatrix} a_{11}, & a_{21}, \dots, & a_{K1} \\ & \dots & \\ a_{1q}, & a_{2q}, \dots, & a_{Kq} \\ & \dots & \\ a_{1Q}, & a_{2Q}, \dots, & a_{KQ} \end{pmatrix}$$

$$\text{bzw.} \qquad \underset{(N\cdot K)}{\mathbf{Z}} = \underset{(N\cdot Q)}{\mathbf{F}} \cdot \underset{(Q\cdot K)}{\mathbf{A}'}$$

Für standardisierte Variablen ist der Korrelationskoeffizient bekanntlich wie folgt definiert:

$$(6) \qquad r_{kj} = 1/N \sum_{i}^{N} z_{ik} z_{ij} \ ; \quad j,k = 1, 2, ..., K$$

Die Korrelationsmatrix

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1K} \\ r_{21} & 1 & \dots & . \\ . & . & . & . \\ r_{K1} & . & . & 1 \end{pmatrix}$$

läßt sich also darstellen als

(7) $R = 1/N\,[Z'Z]$

In diese Gleichung setzen wir nun den in (5) gewonnenen Ausdruck für Z ein:

(8) $R = 1/N(FA')'(FA') - 1/N(AF')(FA') = 1/N(AF'FA')$

Unter der Annahme, daß die Faktoren unkorreliert und ebenfalls z-standardisiert sind, ist die Korrelationsmatrix der Faktoren eine Einheitsmatrix:

(9) $1/N(F'F) = I,$

die in der Hauptdiagonalen lauter Einsen (die Varianzen der Faktoren) enthält. Dies in Gleichung (8) eingesetzt führt zu

(10) $R = AA'$

Den Ausdruck (10) bezeichnet man gelegentlich als »Fundamentaltheorem der deskriptiven Faktoren an alyse« (Hamerle/Kemdny 1981, S. 128).

Für Q=K=3 sehen die Matrizen z. B. so aus:

(10') $$\begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix} = \begin{pmatrix} a_{11}, & a_{12}, & a_{13} \\ a_{21}, & a_{22}, & a_{23} \\ a_{31}, & a_{32}, & a_{33} \end{pmatrix} \cdot \begin{pmatrix} a_{11}, & a_{21}, & a_{31} \\ a_{12}, & a_{22}, & a_{32} \\ a_{13}, & a_{23}, & a_{33} \end{pmatrix}$$

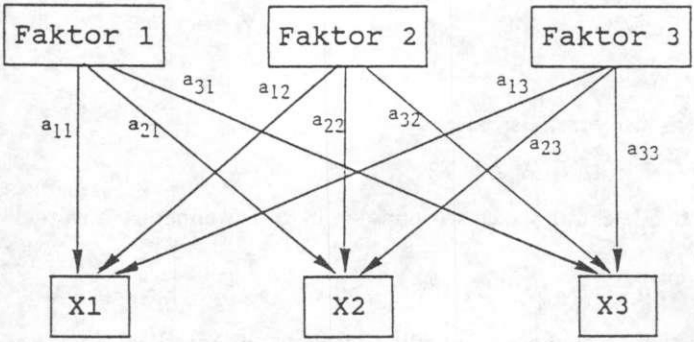
Zur Verdeutlichung: Die Korrelation zwischen z_1 und z_2 ergibt sich aus:

(10'') $r_{12} = a_{11}a_{21} + a_{12}a_{22} + a_{13}a_{23}$

Leser, die mit der Pfadanalyse vertraut sind, können diese Gleichung anhand des folgenden Diagramms (siehe Abb. 2) leicht nach vollziehen:

Die Modellvorgabe, daß die Faktoren sukzessive maximale Varianzanteile Vq (q=l,2,...,Q) erklären sollen, läßt sich hinsichtlich des ersten zu extrahierenden Faktors wie folgt formalisieren:

Abb. 2: Schema der Hauptkomponentenanalyse



$$(11)\qquad V_1 = \sum_{k=1}^{K} (a_{k1})^2 = \max!$$

An der Stelle, an der eine Funktion ein Maximum erreichen soll, muß die erste Ableitung gleich Null sein. Es sollen aber auch noch die Bedingungen (Modellannahmen) gelten, die zur Gleichung (10) führten:

$$(12)\qquad r_{kj} = \sum_{q=1} (a_{kq})(a_{jq}) \;;\; k,j=1,2,...,K$$

$$\Longrightarrow \quad \sum_{q=1} (a_{kq})(a_{jq}) - r_{kj} = 0;\; k,j=1,2,...,K$$

Wenn man berücksichtigt, daß die Korrelationsmatrix symmetrisch ist $(r_{kj}=r_{jk})$ und in der Hauptdiagonalen lauter Einsen enthält $(r_{kk}=r_{jj}=1)$, so sind in Gleichung (12) $q(q-1)/2$ voneinander unabhängige Nebenbedingungen formuliert.

Die zu maximierende Varianz V_1 ist somit eine Funktion von K=Q »Variablen« (den Faktorladungen), die nicht voneinander unabhängig, sondern durch die eben formulierten Nebenbedingungen miteinander verbunden sind. Zur Lösung einer solchen Extremwertaufgabe unter Nebenbedingungen verwendet man in der Regel die Methode der Lagrange'schen Multiplikatoren.[6] Dazu bilden wir eine Hilfsfunktion F, die sowohl die zu maximierende Hauptfunktion (11) als auch die Nebenfunktion (10) bzw. (12) als Komponenten enthält, wobei die Nebenfunktion zusätzlich mit einer unbekannten Konstanten m_{kj} (kj = 1,2,...,K) multipliziert wird:

$$(13)\quad 2F = V_1 - m_{kj}[\sum_{q=1}(a_{kq}\cdot a_{jq}) - r_{kj}]\ ;\ k,j = 1,2,\ldots,K$$

$$2F = (a_{11})^2+\ldots+(a_{K1})^2 + \sum_{k,j=1} m_{kj}r_{kj} - \sum_{k,j=1}\ \sum_{q=1} m_{kj}(a_{kq}\cdot a_{jq})$$

Bei den partiellen Ableitungen nach den einzelnen Faktorladungen a_{kq} tritt der Ausdruck $\Sigma m_{kj}\ r_{kj}$ iimmer nur als Konstante, nicht als Koeffizient der Ladungen auf, so daß wir ihn aus der zu differenzierenden Gleichung eliminieren können:

$$(13^*)\quad 2F = (a_{11})^2 + \ldots + (a_{K1})^2 - \sum_{k,j=1}\ \sum_{q=1} m_{kj}(a_{kq}\cdot a_{jq})$$

In dieser Form ist die Gleichung sowohl in Harman (1967, S. 138) als auch in Arminger (1979, S. 29) angegeben. Implizit wird damit eine (monotone) Gleichungstransformation vorgenommen, die die Lage der gesuchten Extremwerte nicht verändert:

$$(13')\quad F = 1/2\ [(a_{11})^2+\ldots+(a_{K1})^2 - \sum_{k,j=1}\ \sum_{q=1} m_{kj}(a_{kq}\cdot a_{jq})\ ]$$

Die Transformation bietet rechentechnische Vorteile, die hier aber nicht weiter interessieren.[7] Die Konstanten $r_{kj} = m_{jk}$ sind die Lagrange'schen Multiplikatoren (auf die wir später wieder verzichten können). Die K*Q Faktor ladungen a_{kq} sind nun diejenigen »Variablen«, nach denen die Funktion (13') partiell abgeleitet werden muß. (Die einzelnen Ableitungsschritte sind etwas mühsam nachzuvollziehen. Wer sich diese Arbeit ersparen möchte, kann unmittelbar zu dem Ergebnis in Gleichung(21) bzw. (23) vorangehen.) Wir beginnen mit der partiellen Ableitung nach irgendeinem a_{k1}, also der Ladung irgendeiner Z-Variablen auf dem 1. Faktor:

$$(14)\quad \partial F/da_{k1} = a_{k1} - \Sigma m_{kj}a_{j1} = 0$$

Erläuterung: Der Ausdruck a_{k1} stammt aus der Hauptfunktion, der V_1-Komponente in (13'). Da der Faktorindex bei q=1 ebenso wie der Variablenindex k »festgehalten« ist, werden die verschiedenen Koeffizienten, mit denen a_{k1} in den Nebenbedingungen auftritt, nur noch über den Lauf index j = 1, 2..., K generiert.

Als nächstes bilden wir die partielle Ableitung nach einer Ladung a_{kq} der gleichen Z_k-Variablen auf einem anderen Faktor F_q, q#1:

(15) $\partial F / \partial a_{kq} = - \Sigma m_{kj} a_{jq} = 0 ; q \# 1$

Diese Ladungsvariable taucht nur in der Nebenfunktion mit verschiedenen Koeffizienten auf. Im nächsten Schritt fassen wir die Gleichungen (14) und (15) zu einer einzigen Gleichung zusammen. Dazu führen wir ein neues Symbol ein: $\delta_1 q=1$ für q=l und $\delta_1 q=0$ für q#l. Mit dieser Vereinbarung können wir schreiben:

(16) $\partial F / \partial a_{kq} = \partial_{1q} a_{k1} - \Sigma m_{kj} a_{jq} = 0$

Diese Gleichung gilt für jedes beliebige (einzelne) q=l,2,...,Q und für jedes beliebige (einzelne) k=l,2,...,K. Wir benötigen aber die Summe der partiellen Ableitungen nach allen a_{kq} (denn die zu maximierende Varianz ist ja über die Nebenbedingungen eine Funktion aller Ladungen). Die Mathematiker wenden nun rechentechnische »Kniffe« an, deren Sinn vielleicht nicht sofort durchschaubar ist. Sie sind aber harmlos, da sie aus Transformationen bestehen, die die Geltung der Gleichung(en) unberührt lassen. Als erstes wird Gleichung (16) nacheinander mit a_{k1}, k=l,2,...,K multipliziert

(17) $\partial_{1q} (a_{k1})^2 - \Sigma m_{kj} a_{k1} a_{jq} = 0$

Diese K Gleichungen werden sodann summiert:

(18) $\partial_{1q} \sum_{k=1} (a_{k1})^2 \sum_{k=1} \sum_{j=1} m_{kj} a_{k1} a_{jq} = 0 ; q=1,2,...,Q$

Wegen (14) ist

$$\sum_{k=1} m_{kj} a_{k1} = a_{j1}^{(8)}$$

$$\sum_{k=1} m_{kj} a_{k1} = a_{j1}^{(8)}$$

$$\sum_{k=1} (a_{k1})^2 = V_1$$

und wegen (11) ist

$$\sum_{k=1} (a_{k1})^2 = V_1$$

Gleichung (18) läßt sich deshalb schreiben als

(19) $\partial_{1q} V_1 \sum_{j=1} a_{j1} a_{jq} = 0_{kq} a_{jq} = 0 \quad$ für k= 1, 2, ..., K

Diese Gleichung multiplizieren wir mit a_{kq} und summieren über alle q:

(20) $a_{k1} V_1 - \sum_{j=1} a_{j1} \sum_{q=1} a_{kq} a_{jq} = 0 \quad$ für k= 1, 2, ..., K

Erläuterung: Die Summanden $a_{kq} \partial_{1q} V_1$ sind für alle $q \neq 1$ gleich null, da die Größe ∂_{kq} in diesen Fällen vereinbarungsgemäß gleich null ist.

Da auf Grund der vorgegebenen Modellannahmen (Nebenbedingungen) der Ausdruck unter dem zweiten Summenzeichen gleich der Korrelation r_{kj} ist (siehe Gleichung(12)), wird aus(20) nach einer Multiplikation mit (-1)

$$(21) \qquad \sum_{j=1} a_{j1} r_{kj} - V_1 a_{k1} = 0$$

$$\sum_{j=1} r_{kj}\, a_{j1} = V_1\, a_{k1} \qquad \text{für } k=1,2,...,K$$

In Matrizenform läßt sich diese Gleichung wie folgt schreiben:

$$(22) \qquad \begin{pmatrix} r_{12} & \cdots & \cdots & r_{1K} \\ r_{21} & 1 & \cdots & r_{2K} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ r_{K1} & r_{K2} & \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} a_{11} \\ a_{21} \\ \cdot \\ \cdot \\ a_{K1} \end{pmatrix} = V_1 \cdot \begin{pmatrix} a_{11} \\ a_{21} \\ \cdot \\ \cdot \\ a_{K1} \end{pmatrix}$$

In Kurzschreibweise:

$$(23) \qquad Ra_1 = V_1 a_1$$

Die Korrelationskoeffizienten sind bekannt, zu bestimmen ist der Vektor der Faktorladungen (zunächst nur für Faktor 1). Mathematisch ist dieses Problem lösbar mit Hilfe der Theorie der Eigenwerte und der Eigenvektoren. Das Ausgangsproblem wird dort allgemein wie folgt formuliert (siehe z. B. Hamerle/Kemény 1981, S. 187 ff.):

Exkurs: Einige Elemente der Eigen Werttheorie

Gegeben sei eine quadratische Matrix A mit n Zeilen und n Spalten, d. h. mit der Ordnung (n*n). Gesucht wird ein Vektor x (der nicht der Nullvektor ist) und ein Skalar g, die folgende Gleichung erfüllen:

$$(E1) \qquad Ax = gx$$

Man nennt *g* einen *Eigenwert* und x einen *Eigenvektor* der Matrix A. Um einen Lösungsansatz zu erhalten, formt man (E 1) um zu einem homogenen linearen Gleichungssystem:

$$(E\,2) \qquad Ax - gx = 0$$

$$(A - gI_n)x = 0\ , \quad I_n := \text{Einheitsmatrix der Ordnung } n$$

Man nennt (E 2) die *charakteristische Gleichung* der Matrix A. Aus den allgemeinen Bedingungen der Lösbarkeit homogener Gleichungssysteme wissen wir, daß diese Gleichung genau dann eine nicht-triviale Lösung in x hat, wenn die Determinante der Koeffizientenmatrix gleich null ist:

(E3) $|A-gI_n| = 0$

Man bezeichnet den Ausdruck auf der linken Gleichungsseite als *charakteristisches Polynom vom Grade n*, abgekürzt $P_n(g)$. Der Polynomcharakter ergibt sich aus den Regeln der Determinantenentwicklung. Zur Verdeutlichung zitiere ich hierzu ein Beispiel aus Hamerle/Kemény (1981, S. 186 f.):

Gegeben sei die (2*2)-Matrix

$$\mathbf{A} = \begin{pmatrix} 8 & 7 \\ 1 & 2 \end{pmatrix}$$

Das charakteristische Polynom dazu ist

$$\text{(E 4)} \quad |\mathbf{A}-g\mathbf{I}_2| = \left| \begin{pmatrix} 8 & 7 \\ 1 & 2 \end{pmatrix} - g \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = \begin{vmatrix} 8-g & 7 \\ 1 & 2-g \end{vmatrix}$$

$$= (8-g)(2-g) - 7 = g^2 - 10g + 9$$

Dieser quadratische Ausdruck wird gleich null gesetzt und nach *g* aufgelöst:

(E5) $g^2 - 10g + 9 = 0$

$g1 = 9$

$g2 = 1$

Die beiden »Nullstellen« sind die gesuchten Eigenwerte. Allgemein besitzt jedes Polynom n-ten Grades n Nullstellen. Somit gibt es zu einer quadratischen Matrix A der Ordnung (n*n) n Eigenwerte gi (i=l,2,...ji), die nicht notwendigerweise verschieden sein müssen. Die zu g_i gehörenden Eigenvektoren sind Lösungen des Gleichungssystems (E 1). In unserem Beispiel erhält man den Eigenvektor zu dem Eigenwert g_1=9 wie folgt:

(E 6) $$\begin{pmatrix} 8 & 7 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 9 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$8x_1 + 7x_2 = 9x_1 \quad ==> \quad -x_1 + 7x_2 = 0$$
$$x_1 + 2x_2 = 9x_2 \quad ==> \quad x_1 - 7x_2 = 0$$

Der zu g_1 gehörende Lösungvektor ist somit

$$\mathbf{x}_1 = \begin{pmatrix} 7 \\ 1 \end{pmatrix} \text{ bzw. } (\mathbf{x}_1)' = [7 \quad 1]$$

Der zu g_2 gehörende Lösungvektor ergibt sich aus

(E 7) $$\begin{matrix} 7x_1 + 7x_2 = 0 \\ x_1 + x_2 = 0 \end{matrix} \quad ==> \mathbf{x}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

Ein gewisses Problem entsteht dadurch, daß auch alle Vektoren der Form

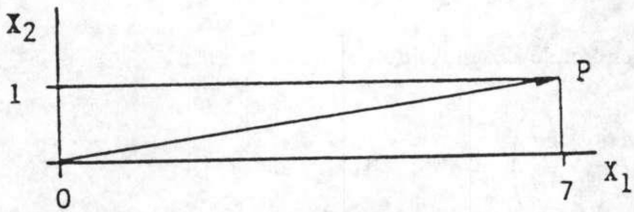
$$\alpha \begin{pmatrix} 7 \\ 1 \end{pmatrix} \text{ und } \alpha \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \alpha \# 0$$

Lösungen des Gleichungssystems (E 1) sind. Man kann jedoch alle Lösungen, die das Vielfache eines bestimmten Vektors darstellen, auf eine bestimmte Form hin gewissermaßen »standardisieren«. Dazu benötigt man das Konzept der »Norm« oder »Länge« eines Vektors x. Sie wird mit $||x||$ symbolisiert und ist wie folgt definiert:

(E8) $$||\mathbf{x}|| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\Sigma x_i^2}$$

Jedes Element des Vektors wird also quadriert; über alle quadrierten Elemente wird die Summe gebildet und daraus die Wurzel gezogen. Der Sinn dieser Operation läßt sich geometrisch deuten. Trägt man die Elemente eines Vektors auf den Achsen eines Cartesischen Kreuzes ein, so schneiden sich die entsprechenden Koordinaten in einem bestimmten Punkt P. Für den ersten Lösungsvektor in unserem Beispiel sieht das wie folgt aus (siehe Abb. 3):

Abb. 3: Geometrische Darstellung
der Länge eines Vektors



Führt man nun einen Pfeil vom Ursprung des Koordinatenkreuzes bis zum Punkt P, so repräsentiert die Länge dieses Pfeils die »Norm« des Vektors *x*\. Das ergibt sich aus dem Lehrsatz des Pythagoras:

(E9) $\quad ||\mathbf{x}|| = \sqrt{(7^2 + 1^2)} = 7{,}071$

Mathematisch läßt sich diese Operation auf Vektoren mit mehr als zwei Elementen (auf ein Koordiantenkreuz mit mehr als zwei Achsen) verallgemeinern.

Dividiert man nun alle Elemente eines Vektors x durch dessen Norm, erhält man einen neuen Vektor $\mathbf{x}^*$, der die Länge 1 hat. In unserem Beispiel mit x' = [7 1] ist

$$\text{(E 10)} \qquad (\mathbf{x}^*)' = \frac{1}{\|\mathbf{x}\|}\,\mathbf{x}' = [(7/7{,}071) \quad (1/7{,}071)]$$

$$\|(\mathbf{x}^*)'\| = \left\|\frac{1}{\|\mathbf{x}\|}\cdot\mathbf{x}\right\| = \sqrt{[(7/7{,}071)^2 + (1/7{,}071)^2]}$$

$$= \sqrt{[(49/50) + (1/50)]}$$

$$= 1$$

Geht man von irgendeinem der anderen Lösungsvektoren $\mathbf{v} = \alpha\mathbf{x}$, $\alpha \# 0$, aus, so läßt sich auch dieser Vektor wieder auf die Länge 1 normieren. Er ist dann identisch mit $\mathbf{x}^*$. Setzen wir z. B. $\alpha=2$, so erhalten wir $\mathbf{v}' = 2(\mathbf{x}')$ = [14 2], $||\mathbf{v}|| = \sqrt{200} = 14{,}142$ und $(\mathbf{v}^*)' = [14/14{,}142 \quad 2/14{,}142] = (\mathbf{x}^*)'$. Die verschiedenen Eigenvektoren, die zu einem bestimmten Eigenwert gehören, bilden jeweils Koordinatenschnittpunkte, die alle auf ein und derselben Geraden liegen, die vom Ursprung des Koordinatenkreuzes ausgeht. (Man trage z. B. den Vektor v in das Koordinatenkreuz der Abb. 3 ein.) Mit anderen Worten, man erhält die verschiedenen zum gleichen Eigenwert gehörenden Eigenvektoren, indem man die Achsen des Koordinatenkreuzes streckt oder staucht.

Die Eigenwerte einer quadratischen Matrix A haben eine Reihe von Eigenschaften, die für verschiedene statistische Verfahren (nicht nur bei der Faktorenanalyse) eine große Rolle spielen. Dazu gehört z. B., daß die Summe der Eigenwerte von A gleich der Summe der Diagonalglieder (also gleich der »Spur«) von A ist, während das Produkt der Eigenwerte gleich der Determinante der Matrix A ist. Wenn also irgendeiner der Eigenwerte g_i (i=1, 2, ..., n) gleich null ist, ist die Matrix A »singulär« und läßt sich nicht »invertieren«.[9] Darüber hinaus führt die Eigenwerttheorie bei symmetrischen Matrizen (wie z. B. einer Korrelationsmatrix) zu besonderen Ergebnissen, die wir bei der Lösung des Faktorenproblems nutzen werden:

(a) Alle Eigenwerte von A sind reell.
(b) Die zu verschiedenen Eigenwerten gehörenden Eigenvektoren sind paarweise orthogonal.
(c) Zu A existiert eine orthogonale Matrix P,[10] mit der man A in folgender Weise zu einer Diagonalmatrix G »zerlegen« kann: P'AP = G bzw. A = PGP'. Dabei sind die Elemente von G, die in der Hauptdiagonalen stehen, die Eigenwerte g_i (i=1,2,...,n) von A und die Spalten vektoren von P bestehen aus paarweise orthonormalen[11] Eigenvektoren von A.

Auf Grund der Eigenschaft (c) läßt sich eine symmetrische Matrix auf folgende Weise in Komponenten zerlegen: Sei p_1 der erste, p_2 der zweite usw. Spaltenvektor von P und g_1 der erste Eigenwert von A, also das erste Element in der Hauptdiagonalen von G, g_2 das zweite usw., so gilt:

(E11) $A = p_1 g_1 p_1' + p_2 g_2 p_2' + \ldots + p_n g_n p_n'$

Es lassen sich auch bestimmte Komponentensets von A abziehen und Restmatrizen A^* bilden, z. B.

(E12) $A^* = A - p_1 g_1 p_1'$

Mit diesem Rüstzeug aus der linearen Algebra können wir zur Lösung des Faktorenproblems, wie es in Gleichung (23) gestellt ist, zurückkehren.

[Ende des Exkurses]

Die Struktur der Gleichung (23) ist identisch mit der Struktur der Eigenwertaufgabe in (E 1). Die Korrelationsmatrix R in Gleichung(23) ist nicht nur eine quadratische, sondern auch eine symmetrische Matrix; folglich kann V_1 als Eigenwert g_1 von R und der Vektor der Faktorladungen a_1 als ein diesem Eigenwert zugeordneter Eigenvektor gelten. Wegen der in Gleichung (11) formulierten Forderung, der 1. Faktor solle einen maximalen Varianzbetrag reproduzieren, wird für $V_1 = g_1$ der maximale Eigenwert herangezogen.[12] Nun sahen wir aber, daß es zu jedem Eigenwert gj einer quadratischen Matrix eine Vielzahl von Eigenvektoren gibt, die alle Lösungen des Gleichungssystems (E 1) darstellen. Das Problem, einen unter den vielen auswählen zu müssen, wird dadurch umgangen, daß man einen beliebigen zu $g_1 = V_1$ gehörenden Eigenvektor herausgreift und, wie oben erläutert, auf die Länge 1 normiert. Diesen Eigenvektor wollen wir hier mit c_1 bezeichnen. Hätten wir irgendeinen anderen zu g_1 gehörenden Eigenvektor ausgewählt und auf die Länge »1« normiert, enthielte er (nach der Normierung) die gleichen Elemente wie c_1. Dieser Vektor ist aber noch nicht der gesuchte Vektor $\mathbf{a_1}$ der Ladungen des 1. Faktors. Dies wird aus folgenden Überlegungen deutlich: Wenn wir nicht nur den ersten, sondern alle K Eigenwerte und die ihnen zugeordneten normierten Eigenvektoren ermitteln, können wir gemäß dem im Exkurs formulierten Zerlegungstheorem die Korrelationsmatrix schrittweise in folgende Komponenten zerlegen (vergl. (E 11)):

(24) $$R = (c_1 g_1 c_1' + c_2 g_2 c_2' + \ldots + c_K g_K c_K')$$

oder in Kurzform:

(25) $R = CGC'$ mit G als Diagonalmatrix der Eigenwerte von R

Diese Darstellung wird erst dann mit dem oben abgeleiteten Fundamentaltheorem der deskriptiven Faktorenanalyse (R = AA') identisch, wenn jede Spalte der C-Matrix $(c_1, c_2, \ldots, c_K)$ mit der Wurzel des entsprechenden Eigenwertes g_i $(i=1,2,\ldots,K)$ gewichtet wird:

(26) $$R = CG^{1/2}G^{1/2}C' = CG^{1/2}(CG^{1/2})'$$ [13]
$$R = AA', \quad A = CG^{1/2}$$

Nun hatten wir allerdings gefordert, die Faktoren sollten sukzessive maximale Varianzanteile »erklären«, der 1. Faktor einen maximalen Anteil der Gesamtvarianz der Z-Variablen, der 2. Faktor einen maximalen Anteil der Restvarianz der Z-Variablen usw. Wir hätten also eigentlich nach Bestimmung des ersten (maximalen) Eigenwertes und des ersten (normierten) Eigenvektors die Rest-Korrelationsmatrix gemäß

(27) $$R^* = R - c_1 g_1 c_1'$$

bilden,[14] sodann für R* wiederum den größten Eigenwert und den zugehörigen Eigenvektor ermitteln, danach eine neue Restmatrix berechnen müssen usw. Dieses umständliche Verfahren ist aber in der Praxis nicht nötig. Es läßt sich nämlich zeigen (siehe Arminger 1979, S. 31), daß der größte Eigenwert von R* der zweitgrößte Eigenwert von R ist und somit unmittelbar den Varianzanteil angibt, den der zweite Faktor erklärt. Entsprechendes gilt auch für die folgenden Eigenwerte der jeweils neu gebildeten Rest-Korrelationsmatrizen.(15) In der Praxis können wir also in einem einzigen Arbeitsgang einfach die Eigenwerte und die jeweils dazu gehörenden Eigenvektoren von R bestimmen, um zu der Darstellung R = CGC' zu gelangen. Die Vektoren der Faktorladungen ergeben sich dann, wie gezeigt, durch die Gewichtung der einzelnen Spalten in C mit der Wurzel des jeweiligen Eigenwertes.(16)

Nachdem die Faktorladungen bestimmt sind, können auch die Faktorwerte (»factor scores«) aus dem Grundmodell (5) der Hauptkomponentenanalyse

(28) $$\mathbf{Z = FA' \text{ bzw. } FA' = Z \Rightarrow F = Z(A')^{-1}}$$

berechnet werden. Das Transponieren und Invertieren der Matrix der Faktorladungen kann man wie folgt umgehen:

$$\mathbf{F = Z(A')^{-1}}$$

(28') $$\begin{aligned}\mathbf{F} &= \mathbf{Z(A')^{-1}}\\ &= \mathbf{Z[(CG^{1/2})']^{-1}}\\ &= \mathbf{Z[G^{1/2}C']^{-1}}\\ &= \mathbf{Z[(C')^{-1}G^{-1/2}]}\\ &= \mathbf{ZCG^{-1/2}}\\ &= \mathbf{Z(AG^{1/2})G^{-1/2}}, \text{ da } \mathbf{A=CG^{1/2}} \text{ laut (26)}\\ &= \mathbf{ZAG^{-1}}\end{aligned}$$

wobei wir die Regel benutzen, daß bei orthogonalen Matrizen - wie C' - die Inverse gleich der Transponierten und die Transponierte einer Diagonalmatrix - wie G - gleich der Ursprungsmatrix ist. Außerdem sei daran erinnert, daß die Inverse der Diagonalmatrix G wiederum eine Diagonalmatrix ist mit den reziproken Eigenwerten als Elementen.(17) Die mit (5) bzw. (28) gegebene Lösung des Faktoren problems ist aber nicht eindeutig. Mit einer orthogonalen (Q*Q)-Matrix T erhält man wegen TT' = I eine zu (28) äquivalente Darstellung (siehe Hamerle/Kernény 1981, S. 193):

(29) $$\begin{aligned}\mathbf{Z} &= \mathbf{FTT'A'}\\ &= \mathbf{F^*A^{*\prime}} \qquad \text{mit } \mathbf{A^* = AT} \text{ und } \mathbf{F^* = FT} \quad \text{(s. Fn. 13)}\end{aligned}$$

Das heißt, man kann die Faktoren und Ladungen orthogonal transformieren, ohne die Grundgleichung (5) bzw. (28) des Hauptkomponentenmodells außer kraft zu setzen. Diese Indeterminiertheit, die zunächst fatal aussehen mag, kann zum Vorteil gewendet werden, indem man eine zunächst erhaltene (beliebige) Faktorenlösung, die vielleicht inhaltlich nicht interpretierbar ist (da sie sich ja nur an dem Kriterium der Maximierung erklärter Varianzen orientiert), so tansformiert, daß eine inhaltlich interpretierbare Lösung erreicht wird. In der Literatur wird dieser Vorgang in der Regel unter dem Titel der »Faktorenrotation« abgehandelt. Sie ist nach der anfänglichen Faktorenextraktion ein zweiter eigenständiger Schritt im Ablauf der Faktorenanalyse. Ich werde ihn aber nicht ausführlich in einem theoretischen Abschnitt behandeln, sondern lediglich im Rahmen des folgenden Analysebeispiels erläutern.

4. Ein Beispiel zur Hauptkomponentenanalyse mit Faktorenrotation

Am Institut für Geographie der Universität Hamburg wurde aus den Volkszählungsakten der Jahre 1840 und 1860 für (alle) 243 Gemeinden des Herzogtums Schleswig eine Vielzahl von demographischen, ökonomischen und sozialen Indikatoren erhoben (siehe Greve 1987; 1988). Für unser Analysebeispiel haben wir aus dem Datensatz für das Jahr 1840 dreizehn Variablen ausgewählt:

Bevölkerungsdichte (BEVDICHT)
Beamte etc. (BEAMTE)
Landwirtschaft. Betriebsleiter (LWBETR)
Angehörige in der Landwirtsch. (LWANG)
Landwirtschaftl. Gesinde (LWGES)
Produzierendes Gewerbe: Meister (PRODMST)
Angehörige im produz. Gewerbe (PRODANG)
Produzierendes Gewerbe: Gehilfen (PRODGEH)
Handel: Selbständige (HANDSELB)
Angehörige im Handel (HANDANG)
Handel: Gehilfen (HANDGEH)
Tagelöhner (TAGL)
Arme (ARME)

Außer der Bevölkerungsdichte sind alle anderen Variablen Anteilsgrößen [18] zur Basis der jeweiligen Einwohnerzahl. Inhaltlich ist dies vielleicht ein wenig aufregendes Beispiel. Man könnte schon aus rein theoretischen Gründen die Beschäftigungsanteile der einzelnen Wirtschaftssektoren (Landwirtschaft, Handel, Gewerbe) jeweils additiv zu Skalen zusammenfassen. Vielleicht möchte man dennoch überprüfen, ob auch eine andere Faktorstruktur vertretbar ist, ob z. B. alle Variablen zu einer einzigen, bipolaren Merkmalsdimension gebündelt werden können. Unklar ist auch, ob die Zahl der Armen, eventuell zusammen mit den Tagelöhnern, einen eigenständigen Faktor bilden oder einem anderen Faktor zuzuordnen sind.

Die ersten Ergebnisse der Faktorenextraktion nach der Hauptkomponentenmethode[19] sind in Abb. 4 a) - d) zusammengefaßt. Dabei wurden die Gemeindedaten mit der jeweiligen Bevölkerungszahl gewichtet (siehe Stahl 1980, S.40).

Abb. 4: Hauptkomponentenanalyse mit 13 Variablen

a) Korrelationsmatrix der Beobachtungsvariablen

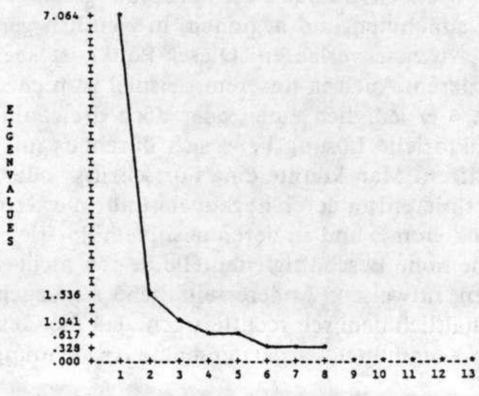
	BEVDICHT	BEAMTE	LWBETR	LWANG	LWGES	PRODMEI	PRODGEH
BEVDICHT	1.00000						
BEAMTE	.29163	1.00000					
LWBETR	-.51899	-.32948	1.00000				
LWANG	-.51091	-.37593	.92509	1.00000			
LWGES	-.44863	-.38327	.67133	.64386	1.00000		
PRODMEI	.47429	.62511	-.42882	-.45419	-.42541	1.00000	
PRODGEH	.62840	.58291	-.61395	-.61913	-.53942	.73720	1.00000
PRODANG	.55008	.63252	-.50862	-.51999	-.47549	.87086	.76944
HANDSELB	.36279	.62789	-.48508	-.48963	-.49736	.72528	.68371
HANDGEH	.40714	.51154	-.48364	-.46874	-.45531	.59303	.74751
HANDANG	.41703	.66845	-.50410	-.51198	-.52923	.72663	.74015
TAGL	-.20677	-.23263	.29688	.30532	.33555	-.17487	-.28532
ARME	.16467	.40401	-.17936	-.21566	-.15350	.23761	.30850

	PRODANG	HANDSELB	HANDGEH	HANDANG	TAGL	ARME
PRODANG	1.00000					
HANDSELB	.75132	1.00000				
HANDGEH	.62122	.75911	1.00000			
HANDANG	.81975	.85810	.70623	1.00000		
TAGL	-.13278	-.23089	-.24625	-.22511	1.00000	
ARME	.30877	.31944	.23021	.39285	.13374	1.00000

b) Anfängliche Faktorextraktion

VARIABLE	COMMUNALITY	*	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
		*				
BEVDICHT	1.00000	*	1	7.06351	54.3	54.3
BEAMTE	1.00000	*	2	1.53650	11.8	66.2
LWBETR	1.00000	*	3	1.04075	8.0	74.2
LWANG	1.00000	*	4	.78545	6.0	80.2
LWGES	1.00000	*	5	.61723	4.7	84.9
PRODMEI	1.00000	*	6	.50450	3.9	88.8
PRODGEH	1.00000	*	7	.41849	3.2	92.0
PRODANG	1.00000	*	8	.32814	2.5	94.6
HANDSELB	1.00000	*	9	.25293	1.9	96.5
HANDGEH	1.00000	*	10	.17495	1.3	97.9
HANDANG	1.00000	*	11	.12382	1.0	98.8
TAGL	1.00000	*	12	.08376	.6	99.5
ARME	1.00000	*	13	.06999	.5	100.0

c) Plot zum Scree-Test



d) Faktorladungen und Xoramunalitäten
bei drei unrotierten Faktoren

	FACTOR 1	FACTOR 2	FACTOR 3	COMMUNALITY
PRODGEH	.89064	.00577	.02211	.79377
HANDANG	.87921	.24674	.09771	.84344
PRODANG	.87276	.23158	.05034	.81788
HANDSELB	.84875	.23469	.17743	.80694
PRODMEI	.82043	.24502	.17113	.76242
HANDGEH	.78611	.10137	.18496	.66246
LWANG	-.74356	.46937	.31231	.87073
LWBETR	-.73501	.50461	.31122	.89173
BEAMTE	.71015	.32865	.13720	.63114
LWGES	-.69100	.39142	.12154	.64547
BEVDICHT	.63199	-.24730	-.18056	.49317
TAGL	-.33068	.51744	-.58168	.71545
ARME	.37471	.48115	-.57815	.70616

Als erstes müssen wir entscheiden, wieviele Faktoren für die weitere Analyse beibehalten werden sollen. Dazu sind eine Reihe von Kriterien vorgeschlagen worden (siehe Kim/Mueller 1978 b, S. 41 ff.). In der Praxis wird am häufigsten das von Kaiser vorgeschlagenes Eigenwertkriterium angewandt: Es werden diejenigen Faktoren beibehalten, die einen Eigenwert von $\geq$ 1.0 aufweisen, die also mindestens soviel Varianz wie eine einzelne Variable aufklären. In unserem Beispiel trifft dies auf drei Faktoren zu. Sie erklären zusammen 74.2 % der Varianz der beobachteten Variablen.

Ein anderes Kriterium zur Bestimmung der Faktorenzahl, der sog. »Scree-Test« ist von Cattell vorgeschlagen worden. Bei diesem Test werden die Eigenwerte (also die erklärte Varianz) auf der Ordinaten, die Rangnummern der Faktoren auf der Abszisse eines Koordinatenkreuzes abgetragen und die Schnittpunkte der Koordinatenpaare durch Linien miteinander verbunden (siehe Abb. 4 c). Die Regel besagt, man solle die Faktorenextraktion an der Stelle beenden, an der diese Verbindungsstücke eine deutlich flachere Neigung annehmen und beginnen, in ziemlich gerader Linie nahezu parallel zur Abszisse verlaufen. Dieser Punkt ist aber oft nicht eindeutig zu identifizieren. Auch in unserem Beispiel ist nicht klar, ob man anhand von Abb. 4 c) lediglich einen oder doch drei Faktoren beibehalten soll. Die einfaktorielle Lösung ließe sich durchaus mit substantiellen Erwägungen stützen. Man könnte eine Fortschritts- oder Modernisierungsdimension definieren, an deren negativem Ende die Gemeinden mit dominanter Agrarökonomie und an deren positivem Pol die Kommunen angesiedelt sind, die hohe Beschäftigtenanteile in den nicht-agrarischen Wirtschaftssektoren aufweisen. Andererseits ließe sich auch die mehrfaktorielle Lösung inhaltlich dadurch rechtfertigen, daß die Gesamtbeschäftigungsquote in den Kommunen variiert und eine Ausdehnung des sekundären und tertiären Sektors nicht sofort zu einer Schrumpfung des primären Sektors führen muß. Wir wollen hier zunächst die dreifaktorielle Lösung weiter verfolgen und später zu der einfaktoriellen zurückkehren.

Die Ladungsmatrix der unrotierten Faktorlösung zeigt mehrere Variablen, die auf mindestens zwei Faktoren gleichzeitig relativ hoch (positiv oder negativ) laden. Zwar zeichnet sich hier schon die Gegenüberstellung von agrarischen und nicht-agrarischen Beschäftigungsindikatoren ab, doch sollte über eine geeignete Transformation der anfänglich extrahierten Faktoren eine eindeutigere Struktur erreichbar sein.

Für die Faktorentransformation (die sog. Rotation)[20] stehen verschiedene formale Kriterien zur Verfügung, die auf leicht unterschiedliche Weise jeweils das Ziel interpretieren, eine sog. »Einfachstruktur« zu erreichen (siehe Arminger 1979, S. 81 im Anschluß an Thurstone). Im Kern geht es darum, Variablengruppen zu finden, die jeweils »stark« auf einem Faktor laden und auf allen anderen nicht bzw. deutlich schwächer. Je eindeutiger dieses Ziel erreicht wird, desto eher lassen sich die errechneten Faktoren als latente Dimensionen interpretieren, denen eine spezifische inhaltliche Bedeutung zukommt. Bei den rechtwinkligen (orthogonalen) Rotationen sind das »Varimax-« und das »Quartimax»-Kriterium am gebräuchlichsten. Ich möchte ihre formalen Eigenschaften hier nicht erläutern; im Ergebnis führen sie nur selten zu größeren Unterschieden (siehe hierzu Arminger 1979, S. 94 f.). Die folgende Abbildung 5 zeigt die Matrix der Faktor ladungen nach einer Quartimax-Rotation.

Abb. 5: Faktorladungen bei QUARTIMAX-Rotation

	FACTOR 1	FACTOR 2	FACTOR 3
HANDANG	.91388	-.07369	.05322
HANDSELB	.89763	-.02920	-.01879
PRODANG	.89396	-.10839	-.08345
PRODMEI	.87301	-.01455	-.00746
PRODGEH	.83771	-.30248	-.02272
HANDGEH	.80086	-.10461	-.10068
BEAMTE	.78827	.07060	.06918
LWBETR	-.47417	.81566	.03990
LWANG	-.49245	.79238	.01892
LWGES	-.50439	.61161	.13034
BEVDICHT	.48059	-.51205	.00103
TAGL	-.26956	.20149	.77601
ARME	.37895	-.07680	.74610

Die Ladungen aller Indikatoren sind nun klar faktorspezifisch mit Ausnahme der Bevölkerungsdichte. Dem ersten Faktor sind diejenigen Indikatoren zugeordnet, die das Gewicht des sekundären und des tertiären Sektors repräsentieren, während der zweite Faktor den Stärkegrad des primären Sektors, der argrar ökonomischen Strukturkomponente, widerspiegelt. Arme und Tagelöhner bilden einen besonderen Faktor, obwohl sie bivariat nur schwach miteinander korrelieren (siehe die Korrelationsmatrix in Abb. 4 a). »Arme« korrelieren positiv mit den Beschäftigtenzahlen des sekundären und des tertiären Sektors - allerdings deutlich schwächer als diese Indikatoren untereinander. Ähnlich ist diese Konstellation bei den Tagelöhnern, die positiv mit den Agrarvariablen korrelieren, aber wiederum schwächer als diese untereinander. Beide Einzelvariablen fallen bei einer rein varianzmaximierenden Faktorenextraktion, bildlich gesprochen, »zwischen« die beiden Variablengruppen und werden auf diese Weise aus ihren angestammten Faktoren sozusagen hinausgedrängt - mit dem Ergebnis, daß sie einen dritten Faktor bilden, der in diesem Falle sogar inhaltlich (als »Armutsfaktor«) interpretierbar scheint.[21] Als störend mag man die Rolle der Variablen »Bevölkerungsdichte empfinden, die den zweiten Faktor bipolar aufspannt, statt mit einem hohen Wert ausschließlich auf dem ersten Faktor positiv zu laden.[22] Orthogonale Faktoren sind in unserem Beispiel inhaltlich wenig sinnvoll, da die Beschäftigten zahlen in der Landwirtschaft mit den Beschäftigtenzahlen in den anderen Sektoren stark negativ korrelieren (siehe Abb. 4 a). Es empfiehlt sich also eine schiefwinklige Faktorentransformation, die man auch als oblique Rotation bezeichnet. (Sie bietet sich zudem immer dann an, wenn eine rechtwinklige Rotation nicht zu der gewünschten Einfach struktur führt.)

Verfahrenstechnischer Ausgangspunkt ist wiederum Gleichung (29). Man verlangt nun nicht mehr, daß die Transformationsmatrix T orthogonal mit TT' = I sein soll. Aus Gleichung (29) wird also

(30)
$$\begin{aligned} \mathbf{Z} &= \mathbf{FTT^{-1}A'} \\ &= \mathbf{F^*(A^*)'} \qquad \text{mit } \mathbf{F^* = FT} \text{ und } \mathbf{A^* = A(T^{-1})'} \end{aligned}$$

Das Ziel ist aber das gleiche wie bei der orthogonalen Transformation: Es soll möglichst eine Ein fach struktur der Faktoren und Variablen erreicht werden. Zur Auswahl einer dazu geeigneten Transformationsmatrix T sind auch hier wieder verschiedene formale Kriterien vorgeschlagen worden (siehe Harman 1967, S. 314 ff.)[23] Das Programmpaket SPSS[x] z. B. bietet die sog. direkte Oblimin-Methode an (s. Arminger 1979, S. 104 - 107).

Als Ergebnis der schiefwinkligen Rotation können die Faktorladungen in Einzelfällen betragsmäßig größer als eins sein. Wenn die Faktoren miteinander korrelieren, ist die Matrix der Faktor ladungen (die man in diesem Zusammenhang auch als »Pattern Matrix« bezeichnet), nicht mehr identisch mit der sog. Strukturmatrix (»Structure Matrix«)die die bivariaten Korrelationen zwischen den Faktoren und den Variablen wiedergibt.

Abb, 6: Faktorladungen und bivariate Korrelationen bei schiefwinkliger Rotation

PATTERN MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3
HANDSELB	.90814	.01435	.03143
PRODMEI	.89203	.03026	.02063
SELBANG	.88511	-.04854	-.04431
BEAMTE	.85753	.12484	-.05212
PRODANG	.83629	-.09661	-.07770
HANDGEH	.76275	-.07733	.10725
PRODGEH	.64994	-.33314	.01481
LWBETR	.09314	.99837	.01055
LWANG	.05977	.97025	.02980
LWGES	-.09475	.72977	-.09604
BEVDICHT	.12764	-.61952	-.03004
TAGL	-.21722	.15669	-.77458
ARME	.25549	-.15553	-.75382

Abb. 7: OBLIMIN-Lösung nach Elimination der Variablen "Bevölkerungsdichte"

PATTERN MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3
PRODMEI	.92710	.08560	.03604
HANDSELB	.89365	-.00765	.01989
HANDANG	.88182	-.05512	-.05260
PRODANG	.88036	-.03109	-.06210
BEAMTE	.83489	.09932	-.06738
HANDGEH	.76857	-.07135	.10373
PRODGEH	.70430	-.26127	.02779
LWBETR	.05696	.99271	.02074
LWANG	.02681	.96817	.04114
LWGES	-.11050	.74002	-.08052
ARME	.24140	-.17865	-.77136
TAGL	-.20438	.18112	-.75786

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3
HANDANG	.91684	-.56403	-.09562
PRODANG	.90189	-.53821	-.10741
HANDSELB	.89701	-.53059	-.02845
PRODMEI	.87521	-.45842	-.02332
PRODGEH	.85509	-.67460	.01481
HANDGEH	.80447	-.52962	.06852
BEAMTE	.78068	-.38077	-.12302
LWBETR	-.52291	.95746	-.08012
LWANG	-.53988	.94849	-.05564
LWGES	-.53750	.81237	-.14730
ARME	.38793	-.24345	-.76703
TAGL	-.26833	.37488	-.76446

FACTOR CORRELATION MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3
FACTOR 1	1.00000		
FACTOR 2	-.58298	1.00000	
FACTOR 3	-.05494	-.09844	1.00000

Die Abbildung 6 zeigt Ergebnisse einer schiefwinkligen Rotation nach dem OBLIMIN-Kriterium. Die beiden ersten Faktoren korrelieren negativ miteinander (r = -.60), was den inhaltlichen Überlegungen entspricht.

Wenn wir die etwas problematische Bevölkerungsvariable aus der Korrelation smatr ix eliminieren, erhalten wir nach dem Kaiser-Kriterium wiederum drei relevante Faktoren, die zusammen 76.9% der Varianz erklären. Die Ladungsmatrix der schiefwinkligen Rotation ändert sich für die übrigen Variablen aber nur geringfügig (siehe Abb. 7). Die Korrelation zwischen den beiden ersten Faktoren beträgt nun r = -.583.[24]

Die »Pattern Matrix« läßt erkennen, aus welchen Variablen die verschiedenen Faktoren primär gebildet werden. Faktor 1 z. B. wird fast ausschließlich aus den Beschäftigungsindikatoren des sekundären und des tertiären Sektors konstruiert. Die »Structure Matrix« deckt anhand der bivariaten Korrelationen zwischen den Faktoren und den Beobachtungsvariablen zusätzliche Bedeutungsgehalte auf, die in einem Faktor impliziert sein können. Bei Faktor 1 ist es z. B. die tendenzielle aber nicht perfekte negative Korrelation zwischen den Beschäftigungsniveaus des agrarischen und des nicht-agrarischen Bereichs. (Deutliche negative Koeffizienten für die Agrarindikatoren treten in der Structure, nicht aber in der Pattern Matrix auf.)

Wir wollen das in Abb. 7 enthaltene Ergebnis auch benutzen, um die Faktorenwerte zu berechnen, also die Positionen der verschiedenen Kommunen auf den latenten Dimensionen zu ermitteln. Standard-Computerprogramme (wie SPSS oder BMDP) berechnen sie nicht direkt über Gleichung (28'), sondern wenden dazu den Algorithmus der Regressionsschätzung an, der unten in Abschn. 5 zur Hauptfaktorenanalyse erläutert wird (siehe Gleichung (31) ff. sowie Fn. 28). Er führt im Falle der Hauptkomponentenmethode zu exakten Ergebnissen, die denen von Gleichung (28) entsprechen. Ein Zwischen schritt ist die Berechnung der sog. Faktor-Beta-Ladungen (»Factor-Score Coefficients«), die stets in den Computer-Ausdrucken erscheinen und ebenfalls in Abschnitt 5 erläutert werden.[25] Abb. 8 listet diese Koeffizienten für alle drei Faktoren. Sie stellen Regressionskoeffizienten dar, mit denen die Z-Variablen gewichtet werden, um die Faktorwerte durch Linearkombination zu »erzeugen«.

Abb. 8: Faktor-Beta-Koeffizienten (BMDP-Ausdruck)

```
FACTOR SCORE COEFFICIENTS
-------------------------
THESE COEFFICIENTS ARE FOR THE STANDARDIZED VARIABLES,
MEAN ZERO AND STANDARD DEVIATION ONE.

                   FACTOR     FACTOR     FACTOR
                      1          2          3

beamte      3    0.16554    0.05256    0.05179
lwbetr      4    0.02903    0.38022   -0.02176
lwang       5    0.02276    0.37040   -0.03849
lwges       6   -0.00886    0.28030    0.06432
prodmei     7    0.18389    0.04941   -0.03460
prodgeh     8    0.13387   -0.08708   -0.02523
prodang     9    0.17217    0.00359    0.04775
handselb   10    0.17558    0.01313   -0.02062
handgeh    11    0.15025   -0.01307   -0.08950
handang    12    0.17207   -0.00552    0.03994
tagl       13   -0.04041    0.06229    0.63059
arme       14    0.04071   -0.06720    0.64132
```

Die Faktorwerte, die wir in unserem Beispiel erhalten (siehe unten Abb. 9), resultieren aus einem Rechnungsgang einer Fallgewichtung anhand der Bevölkerungsgröße der jeweiligen Kommune. Wegen der Fallgewichtung sind die Faktorwerte selbst nicht z-standardisiert; weder ist ihr Mittelwert gleich null, noch ist die Standardabweichung gleich 1 (im Unterschied zu den Faktorwerten, die aus einer ungewichteten Hauptkomponentenanalyse resultieren.) Berechnet man aber in nachfolgenden Analysen diese beiden Statistiken für die entsprechend gewichteten Faktorwerte, erhält man als arithmetisches Mittel Null und als Standardabweichung die Eins. Die gewichteten Korrelationen der Faktorwerte sind auch wieder identisch mit der Matrix der Faktorkorrelationen, die direkt mit Hilfe der Transformationsmatrix errechnet werden (s. Fn. 24).

Die inhaltliche Interpretation einer latenten Dimension, die als »Faktor« dargestellt wird, ergibt sich, wie bereits erwähnt, aus den Variablen, die auf dem betreffenden Faktor (betragsmäßig) hoch laden bzw. mit ihm hoch (positiv oder negativ) korrelieren. Der erste Faktor in unserem Beispiel faßt das Beschäftigungsniveau in den Bereichen Gewerbe, Handel und öffentliche Dienstleistungen zusammen. Wir wollen ihm das Kürzel »Sekundärökonomie« anheften. Der zweite Faktor stellt das Beschäftigungsniveau im Agrarbereich dar; er erhält das Kürzel »Agrarökonomie«. Im Unterschied zu den Ursprungsvariablen (Anteilsgrößen) liegt den Faktorwerten keine inhaltlich interpretierbare Skaleneinheit zugrunde. Dennoch ist die relative Position einer Kommune, die sie im Vergleich zu allen anderen Kommunen auf der latenten Dimension einnimmt, durch die jeweiligen Faktorwerte bestimmbar. Abb. 9 zeigt für beide Faktoren die jeweils fünf Gemeinden mit den höchsten und niedrigsten Skalenwerten.

Abb. 9: Gemeinden mit extremen Faktorwerten

Faktor 1		Faktor 2	
-1.30	Bov	-2.03	Christiansfeld
-1.28	Handewitt	-1.70	Eckernfoerde
-1.26	Bulderup	-1.52	Soenderborg
-1.25	Raepsted	-1.50	Toender
-1.18	Wallsbuell	-1.50	Flensburg
2.20	Husum	2.10	Hollingstedt
2.23	Eckernfoerde	2.28	Enge
2.36	Garding	2.31	Hjerndrup
2.66	Toender	2.41	Hallig Groede
3.27	Friedrichstadt	2.44	Toerstrup

Kehren wir nun noch einmal zu der eingangs angesprochenen einfaktoriellen Lösungsmöglichkeit zurück, dem Versuch, alle Gemeinden auf einem bipolaren Faktor anzuordnen, den man als »Sektorale Beschäftigungsstruktur« bezeichnen und (vielleicht) als Modernisierungsdimension deuten mag. Der negative Pol dieses Faktors repräsentiert das Gewicht des primären, der positive Pol das Gewicht des sekundären und des tertiären Sektors. Wenn wir die beiden Variablen »Tagelöhner« und »Arme« aus der Analyse herausnehmen, erhalten wir eine Ladungsmatrix (siehe Abb. 10), die sich von der ursprünglichen, unrotierten Lösung in Abb 4 d kaum unterscheidet. Wenn wir nur einen bipolaren Faktor mit maximaler Varianz skalieren wollen, entfällt der Rotationsschritt. Die Faktor-Beta-Ladungen (»Factor Score Coefficients«) des 1. Faktors unterscheiden sich von denen der unrotierten Lösung vor allem in den (negativen) Koeffizienten für die Agrarindikatoren. In Abb. 11 sind für beide Pole die Gemeinden mit den extremsten Werten gelistet.

Abb. 10: Ladungen bei unrotierten Faktoren
nach Ausschluß der Variablen
BEVDICHT, TAGL, ARME

		FACTOR 1	FACTOR 2
handang	12	0.890	0.000
prodgeh	8	0.882	0.000
prodang	9	0.879	0.000
handselb	10	0.866	0.000
prodmei	7	0.831	0.296
handgeh	11	0.797	0.000
lwang	5	-0.734	0.599
lwbetr	4	-0.727	0.629
beamte	3	0.716	0.321
lwges	6	-0.685	0.448

Am negativen Ende finden wir Kommunen, die in der zweifaktorieilen Lösung hohe positive Werte im Faktor »Agrarökonomie« aufweisen (vergl. Abb. 9); am positiven Ende stehen jetzt Gemeinden, die zuvor hohe Werte in der Skala »Sekundärökonomie« einnahmen.

Abb. 11: Extremwerte auf bipolarem Faktor
(unrotierte Hauptkomponentenlösung)

-1.63	Hallig Groede
-1.21	Hollingstedt
-1.11	Olderup
-1.10	Sommersted
-1.08	Skrave
2.16	Soenderborg
2.19	Eckernfoerde
2.23	Garding
2.53	Toender
2.98	Friedrichstadt

5. Die Hauptfaktorenmethode[26]

Die Hauptkomponentenmethode ist durch die Vorgabe gekennzeichnet, daß es nur gemeinsame Faktoren gibt, die die gesamte Varianz der beobachteten Variablen erklären. Die Hauptfaktorenmethode läßt diese Voraussetzung fallen; sie sieht statt dessen vor, daß neben den gemeinsamen auch variablenspezifische Faktoren (oder allgemeiner: »Einflußgrößen«) wirksam sind, daß also die Anzahl der gemeinsamen Faktoren F_q (q=1,2,...,Q) kleiner ist als die Anzahl der beobachtete Variablen X_k (k=1,2,...,K): Q<K. Somit kehren wir zu dem Ausgangsmodell in Gleichung (3) und Abb. 1 zurück:

$$(3') \qquad Z_k = a_{k1}f_1 + a_{k2}f_2 + \ldots + a_{kQ}f_Q + u_k \; ; \; k=1,2,\ldots,K$$

Die Korrelationsmatrix der X-Variablen läßt sich nun nicht mehr exakt aus den quadrierten Ladungen der gemeinsamen Faktoren rekonstruieren; aus Gleichung (10) wird

$$(30) \qquad \mathbf{R = AA' + U^2}$$

(Wir gehen weiterhin von z-standardisierten Variablen aus.) Die Matrix U^2 enthält für jede Variable Z_k denjenigen Varianzanteil, der nicht durch die gemeinsamen Faktoren erklärt werden kann. Da vorausgesetzt wird, daß die Einzelrestfaktoren der verschiedenen Z-Variablen nicht miteinander korrelieren, ergibt $U'U = U^2$ eine Diagonalmatrix. Die quadratische Matrix $[R - U^2]$ unterscheidet sich (im theoretischen Modell) von R also nur in den Diagonalelementen: In der Hauptdiagonalen stehen nicht mehr lauter Einsen (also die Varianzen der Z-Variablen), sondern die Kommunalitäten $(h_k)^2 = 1 - (u_{kg})^2$, also diejenigen Varianzanteile der Z-Variablen, die durch die gemeinsamen Faktoren erklärt werden (und die im Hauptkomponentenmodell mit Q=K für alle Variablen gleich eins waren). U^2 und damit die Kommunalitäten sind nicht bekannt und werden auch nicht auf dem Wege vorgegebener Modellrestriktionen festlegt. Wenn man sie aber in irgendeiner Weise schätzen könnte, ließe sich das aus der Hauptkomponentenmethode bekannte Verfahren der Faktorenextraktion (Zerlegung der Korrelationsmatrix) statt auf die Korrelationsmatrix R auf die Matrix $[R - U^2]$ anwenden.

Üblicherweise werden die Kommunalitäten in einem iterativen Verfahren geschätzt. Als Anfangsschätzer für $(h_k)^2$ setzt man in der Regel den quadrierten multiplen Korrelationskoeffizienten $(r_{k.})^2$ ein, der sich als Bestimmtheitsmaß einer Regression der Variablen X_k auf alle anderen X-Variablen errechnen läßt. Es gilt nämlich $(r_{k.})^2 \leq (h_k)^2$, k=1, 2, ..., K (zum Beweis siehe Arminger 1979, S. 42 f.). Da dieser Schätzer somit eine Mindestgröße für $(h_k)^2$ liefert, stellt die Anzahl der positiven Eigenwerte der Matrix $[R - U^2]$ mit $(u_k)^2 = 1 - (r_{k.})^2$ ein Minimum für die Anzahl der gemeinsamen Faktoren dar, die überhaupt einen Anteil zur Erklärung der Gesamtvarianz leisten.[27] In der Hauptfaktorenanalyse wird deshalb die Zahl der positiven Eigenwerte häufig als Kriterium (sog. »Guttman-Kriterium«) für die Anzahl der zu extrahierenden Faktoren herangezogen. Neben inhaltlichen Gesichtspunkten sind auch weitere formale Auswahlkriterien vorgeschlagen worden (siehe Kim/Mueller 1978, S. 43 f.), z. B.: die Faktorenzahl in einer vorgeschalteten Hauptkomponentenanalyse nach dem bereits genannten Kaiser-Kriterium zu bestimmen (Arminger 1979, S. 42). Hat man die Anzahl der Faktoren anhand irgendeines Kriteriums bestimmt, läßt sich der Kommunalitätenschätzer iterativ verbessern: Mit dem in Abschnitt 3 dargestellten Hauptkomponentenverfahren berechnet man die Eigenwerte und einen ersten Satz von Faktor ladungen durch Zerlegung der Matrix $[R - U^2]$ mit $(u_k)^2 = 1 - (r_k.)^2$. Mit den Summen der quadrierten Faktorladungen jeder Variablen erhält man einen neuen Satz geschätzter Kommunalitäten. Diese neuen Kommunalitätenschätzer setzt man in die Hauptdiagonale der Korrelationsmatrix ein und wiederholt den Vorgang der Faktorenextraktion bis das Verfahren konvergiert, bis sich also keine nennenswerten Änderungen der Kommunalitätenschätzer ergeben.

Wenden wir die Hauptfaktorenanalyse auf unseren Beispieldatensatz aus Abschnitt 4 an, erhalten wir (bei Auswahl der Faktoren nach dem Kaiser-Kriterium) folgendes Ergebnis für die Schätzung der Kommunalitäten (siehe Abb. 12)

Abb. 12: Ergebnis der iterierten Kommunalitätenschätzung

VARIABLE	COMMUNALITY	*	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
		*				
BEAMTE	.53269	*	1	6.44513	53.7	53.7
LWBETR	.94767	*	2	1.24586	10.4	64.1
LWANG	.87947	*	3	.78596	6.5	70.6
LWGES	.52305	*				
PRODMEI	.73452	*				
PRODGEH	.75301	*				
PRODANG	.79967	*				
HANDSELB	.77926	*				
HANDGEH	.60200	*				
HANDANG	.83763	*				
TAGL	.21543	*				
ARME	.87256	*				

Der dritte Faktor hat nun einen Eigenwert kleiner 1. In der Hauptfaktorenanalyse ist die Summe der Eigenwerte (d. h. auch die Summe der Kommun alitäten) bei gleicher Faktorenzahl kleiner als in der Hauptkomponentenlösung. In unserem Beispiel sind zudem die Kommunalitäten in der Hauptfaktorenanalyse gegenüber der Hauptkomponentenlösung etwas verschoben, insbesondere weist die Variable »Tagelöhner« einen extrem niedrigen Wert auf. Die Oblimin-Rotation führt zu einer ähnlichen Faktorstruktur wie die schiefwinklige Rotation in der Hauptkomponentenanalyse (siehe Abb. 13 im Vergl. zu Abb. 7); Faktor 3 wird nun aber in den sehr ungleichen Ladungsgewichten für »Tagelöhner« und »Arme« deutlicher als Artefakt erkennbar. Wenn wir diese beiden Variablen aus der Analyse ausschließen, erhalten wir eine zweifaktorielle Lösung, deren Ladungsmatrix nach einer OBLIMIN-Rotation den Ladungen der restlichen Variablen auf den beiden ersten Faktoren in Abb. 13 sehr ähnlich ist.

Im Unterschied zur Hauptkomponentenanalyse können in der Hauptfaktorenanalyse die Faktorwerte nicht direkt aus den z-Werten der Beobachtungsvariablen, den Faktorladungen und Eigenwerten ermittelt werden (siehe Gleichung (28')). Sie sind mit Hilfe der Regressionsrechnung zu bestimmen (siehe Arminger 1979, S. 115 f.).

Jeder Faktor F_q ist als Linearkombination der entsprechend gewichteten Beobachtungsvariablen Z_k plus einer Restgröße e_q darstellbar:

(31) $$\underset{N\cdot Q}{\mathbf{F}} = \underset{N\cdot K}{\mathbf{Z}} \cdot \underset{K\cdot Q}{\mathbf{B}} + \underset{N\cdot Q}{\mathbf{E}},$$

wobei B die Matrix der Gewichte (Regressionskoeffizienten) darstellt.

Wären die Faktorwerte bekannt, ließe sich B nach dem üblichen Kleinstquadratverfahren schätzen:

(32) $$B = (Z'Z)^{-1}(Z'F)$$

Obwohl F nicht bekannt ist, läßt sich mit Hilfe von Modellrestriktionen eine Lösung entwickeln: Wenn wir in Zähler und Nenner den Faktor 1/N einfügen, erhalten wir die äquivalente Darstellung

(33) $$B = [(1/N)Z'Z]^{-1}[(1/N)Z'F]$$

Da wir von z-standardisierten Variablen ausgegangen sind, stellt der erste Ausdruck in eckigen Klammern deren Korrelationsmatrix R dar. Wenn wir voraussetzen, daß auch die Faktoren z-standardisiert sind, stellt der zweite Ausdruck in eckigen Klammern die Matrix der Korrelationen zwischen den Faktoren und den beobachteten, z-transformierten Variablen dar. Wir wollen dafür das Symbol Weinführen. (Dabei handelt es sich um die bereits bekannte »Strukturmatrix«.) Solange die Faktoren orthogonal zueinander stehen (unkorreliert sind), ist die Strukturmatrix identisch mit der Matrix A der Faktorladungen. Wir können also schreiben

(33') $$B = R^{-1}W = R^{-1}A^{(28)}$$

Damit sind die Regressionskoeffizienten (»Faktor-Beta-Ladungen«, »factor score coefficients«) im Gleichungssystem (31) auf bekannte Größen

ABB. 13: Faktorladungen und bivariate Korrelationen nach OBLIMIN-Rotation

PATTERN MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3
PRODMEI	.91648	.09539	.06166
PRODANG	.90086	.01350	-.01532
HANDANG	.89459	-.01888	-.07105
HANDSELB	.88928	.01007	.00458
BEAMTE	.72615	.03355	-.12745
HANDSELB	.72540	-.08170	.06187
PRODGEH	.70761	-.23093	.00851
LWBETR	.07313	1.01988	.03422
LWANG	.02032	.95485	.06147
LWGES	-.21010	.57215	-.03865
ARME	.18282	-.16335	-.88168
TAGL	-.13952	.23715	-.30278

STRUCTURE MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3
HANDANG	.91260	-.55758	-.15057
PRODANG	.89401	-.53497	-.09886
HANDSELB	.88272	-.53346	-.07754
PRODMEI	.85261	-.47085	-.03205
PRODGEH	.84785	-.66394	-.03132
HANDGEH	.76966	-.53128	.00448
BEAMTE	.71728	-.39626	-.19722
LWBETR	-.55278	.97157	-.08135
LWANG	-.56836	.93588	-.04234
LWGES	-.55597	.70458	-.08059
TAGL	-.25674	.35468	-.31539
ARME	.36293	-.18085	-.88090

FACTOR CORRELATION MATRIX:

	FACTOR 1	FACTOR 2	FACTOR 3
FACTOR 1	1.00000		
FACTOR 2	-.61065	1.00000	
FACTOR 3	-.09114	-.10678	1.00000

zurückgeführt. Mit ihrer Hilfe lassen sich nun auch die Faktorwerte gemäß (31) schätzen (s. Norusis 1985, S. 148). Im Ergebnis zeigen sich nur geringe Unterschiede im Vergleich zu den Skalen, die wir in Abschnitt 4 vorgestellt haben.

6. Schlußbemerkungen

In diesem Papier haben wir die Modellannahmen und die mathematischen Operationen erläutert, mit denen in der »Hauptkomponentenanalyse« und der »Hauptfaktorenanalyse« latente Dimensionen, »Faktoren«, aus beobachteten Variablen konstruiert werden können. Es gibt noch eine Reihe anderer Methoden der Faktorenextraktion, wie z. B. die Methode der Kleinsten Quadrate,[29] die Maximum Likelihood-Methode und die sog. Alpha-Methode, die wir hier nicht behandeln wollen (siehe z. B. Arminger 1979, Holm 1976, Kim/Mueller 1978, Bacher 1990). Alle diese Methoden bieten gewisse Variationsmöglichkeiten in den allgemeinen Modellannahmen sowie unterschiedliche Möglichkeiten der statistischen Kontrolle der Ergebnisse.

Hauptkomponentenanalyse und Hauptfaktorenanalyse habe ich hier lediglich als explorative Verfahren behandelt. In diesem Falle sind keine Annahmen über spezifische Verteilungseigenschaften (z. B. Vorliegen einer Normalverteilung) der beobachteten Variablen nötig (siehe Holm 1976, S. 59; Arminger 1979, S. 24). Sie ermöglichen es dem Forscher, »datenorientiert« zu arbeiten, auch wenn er keine präzisen Hypothesen zur Faktorstruktur (Zahl der Faktoren und Zuordnung der Indikatoren zu spezifischen Faktoren) hat. Allerdings führen die explorativen Verfahren häufig zu mehreren Lösungen, die plausibel oder vertretbar scheinen; in ihrem Rahmen kann jedoch nicht statistisch getestet werden, welches Modell den vorliegenden Daten am besten entspricht. Will man die Adäquanz spezieller Modelle testen, muß man zur sog. konfirmatorischen (hypothesentestenden) Faktorenanalyse übergehen (siehe Weede/Jagodzinski 1977; Arminger 1979; Bacher 1990, Kap. 5). Sie setzt (in ihren verschiedenen Varianten) neben einer präzisen Theorie auch spezifische Verteilungseigenschaften der Variablen voraus. Dafür bietet sie eindeutigere und statistisch abgesicherte Entscheidungshilfen.

Einige der Arbeitsschritte, die auch in der explorativen Faktorenanalyse schon möglich oder nötig sind, sind hier nicht dargestellt worden. Dazu gehört die Überprüfung der Homogenität oder Heterogenität der Untersuchungseinheiten (die Grundgesamtheit kann sich aus verschiedenen Subpopulationen zusammensetzen, für die unterschiedliche Faktorstrukturen gelten). Heterogenität der Population sowie »Ausreißer« (die z. B. in Streudiagrammen der Faktorskalen sichtbar werden) können die Faktorstruktur erheblich verzerren (siehe hierzu Bacher 1990, S. 127 f., 136 f.). Auch wenn die Faktorenanalyse nur explorativ eingesetzt wird, sollte man (falls die Anwendungsvoraussetzungen erfüllt sind) bestimmte Tests durchführen, mit deren Ergebnis man besser einschätzen kann, ob bei gegebenener Korrelationsmatrix eine Faktoren analyse überhaupt sinnvoll ist (siehe Norusis 1985, S. 127 ff.)

Literaturverzeichnis

Arminger, Gerhard, Faktorenanalyse. Stuttgart 1979.

Bacher, Johann, Skalierungsverfahren, Historical Social Research/Historische Sozialforschung, Vol 15, No. 3 (Special Issue), Köln 1990.

Bortz, Jürgen, Lehrbuch der Statistik. Berlin, Heidelberg, New York 1979.

Draper, N. R. und Smith, H.: Applied Regression Analysis, New York u.a. 1981[2].

Fox, John: Describing Univariate Distributions, in: J. Fox und J. Scott Long (eds.), Modern Methods of Data Analysis, Newbury Park 1990, S. 58 - 125.

Greve, Klaus, Zentrale Orte im Herzogtum Schleswig 1860. Studien zur Wirtschafts- und Sozialgeschichte Schleswig-Holsteins, Band 12. Neumünster 1987.

Greve, Klaus, Volkszählungen und Landgewerbelisten in Schleswig-Holstein in dänischer Zeit - Quellenmaterial zur Berufsstatistik, in: Ingwer E. Momsen (Hrsg.), Schleswig-Holsteins Weg in die Moderne. Zehn Jahre Arbeitskreis für Wirtschafts- und Sozialgeschichte Schleswig-Holsteins. Neumünster 1988.

Hamerle, Alfred/Kem^ny, Peter, Einführung in die Mathematik für Sozialwissenschaftler. München und Wien 1981.

Harman, Harry H., Modern Factor Analysis. Chicago, London 1967[2].

Holm, Kurt, Die Befragung 3. Die Faktorenanalyse. München 1976.

Kim, Jae-On/ Mueller, Charles W., Introduction to Factor Analysis. What It Is and How To Do It. Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills, London 1978 a.

Kim, Jae-On/ Mueller, Charles W., Factor Analysis. Statistical Methods and Practical Issues. Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills, London 1978 b.

Linder, Arthur und Berchthold, Willi, Statistsiche Auswertung von Prozentzahlen, Basel und Stuttgart 1976.

Nomsis, Marija J., SPSS[x] Advanced Statistics Guide. New York u. a. 1985.

Stahl, Herbert, Beschreibung der Sozialstruktur in Berlin (West) mit Hilfe der Faktorenanalyse, in: Berliner Statistik, Heft 3, 1980.

Weede, Erich/ Jagodzinski, Wolfgang, Einführung in die konfirmatorische Faktorenanalyse, in: Zeitschrift für Soziologie 6 (1977), S. 315 - 353.

Überla, K., Faktorenanalyse. Eine systematische Einführung für Psychologen, Mediziner, Wirtschafts- und Sozialwissenschaftler. Berlin, Heidelberg, New York 1968.

Notes

(1) Für Ratschläge und praktische Unterstützung danke ich meinen Kollegen Steffen Kühnel und Ralph Ponemereo. Danken möchte ich auch Herrn Dr. Klaus Greve, Universität Osnabrück, der die Daten für die Beispielrechnungen zur Verfügung stellte.

(2) Zur Faktoren an alyse mit nicht-metrischen Variablen siehe Arm in - ger (1979).

(3) Aber auch in diesem Fall kann man die Beziehung zwischen Konstrukt und Indikator »technisch« als eine Relation von Ursache und Wirkung behandeln.

(4) Solche Interpretationen findet man gelegentlich in der Literatur. Allerdings ist unklar, was eine (wie auch immer reduzierte) »Information« sein soll, wenn sie nicht (im Sinne einer latenten Merkmalsdimension) theoretisch deutbar ist.

(5) Inhaltlich ist dies in vielen Fällen keine sinnvolle Annahme. Deshalb wird in späteren Schritten (siehe Ab sehn. 4) oft eine Transformation auf korrelierte Faktoren durchgeführt (Arminger 1979, S. 18).

(6) Siehe z. B. die Kurzdarstellung in Van de Geer (1971: 71 ff.). Die folgende Darstellung ist eine detailliertere Version der entsprechenden Ableitungen in Harman (1967, S. 138 f.) und Arm in ger (1979, S. 29 f.).

(7) Indem man die rechte Gleichungsseite durch 2 dividiert, erhält man bei der Ableitung nach einem quadrierten Term, z. B. $(a_{11})^2$, nicht $2a_{11}$, sondern a_{11} als Ergebnis. Ebenso führt die Ableitung von $2a_{31}a_{11}$ nicht zu $2a_{31}$, sondern zu a_{31}. Eine solche monotone Transformation verändert nicht die Lage der Maxima und Minima.

(8) Gegenüber Gleichung (14) sind (aus zunächst vielleicht nicht einsehbaren Gründen) die Indizes k und j, die den Ladungen zugeordnet sind, vertauscht. Inhaltlich wird aber in beiden Fällen das gleiche ausgesagt: Die Ladung irgendeiner Variablen auf dem 1. Faktor ist identisch mit der Summe der Ladungen aller Variablen auf diesem Faktor, sofern sie mit geeigneten (zunächst noch unbekannten) Multiplikatoren gewichtet werden.

(9) Eine Matrix, die nur positive Eigenwerte aufweist, nennt man »positiv definit«. Eine Matrix, die keine negativen, aber nicht ausschließlich positive Eigenwerte hat, nennt man »singulär«.

(10) Eine quadratische orthogonale Matrix P ist durch die Eigenschaft definiert, daß P'P = PP' =I.

(11) Zwei Vektoren a und b heißen »orthonormal«, wenn das Produkt a'b = 0 ist und sie auf die Länge »1« normiert worden sind.

(12) In den Standardcomputerprogrammen wird neben dem Betrag der erklärten Varianz (also dem Eigenwert) auch der Anteil der erklärten Varianz, g_q/Q, ausgewiesen.

(13) Allgemein gilt: (AB)' = B'A' sowie: die Transponierte einer Diagonalmatrix (wie G) ist identisch mit der Ursprungsmatrix.

(14) Das heißt, für jedes Element r(kj) der Korrelationsmatrix R wäre auf diese Weise ein Element

$$r(kj)^* = r(kj) - c_{k1}g_1(c_{j1}) = r(kj) - a_{k1}a_{j1}$$
$$= a_{k2}a_{j2} + a_{k3}a_{j3} + \ldots + a_{kQ}a_{jQ}$$

für die Rest-Korrelationsmatrix R* gebildet worden. Zur Verdeutlichung siehe Gleichung (10'') mit k=l und j=2. Von r(kj) wird derjenige Anteil subtrahiert, der durch das Produkt der Ladungen a_k und aj, (kj=l,2,...,K; j#k) auf (nur) dem ersten Faktor gegeben ist (siehe auch Abb. 3).

(15) Der Eigenvektor, der dem größten Eigenwert (gj) in R entspricht ist auch ein Eigenvektor in R*, aber der mit ihm verbundene Eigenwert in R* ist gleich null (siehe Harman 1967, S. 142).

(16) Mit der Normierung und Gewichtung der Vektoren sind die Skaleneinheiten der Faktoren festgelegt. Das kann man sich anhand der Regressionsanalyse verdeutlichen, wo man den umgekehrten Weg geht. Skaliert man z. B. eine Regressorvariable nicht in DM-, sondern in Pfennigeinheiten, führt das zu einer Verringerung des Regressionsgewichts (Steigungskoeffizienten) um den Faktor 100.

(17) Ahand von (28') läßt sich auch bestätigen, daß die Varianz/Kovarianz-Matrix der Faktoren eine Einheitsmatrix darstellt:

$$\mathbf{F} = \mathbf{ZAG^{-1}}$$
$$\mathbf{1/N)F'F} = (1/N)(\mathbf{ZAG^{-1}})'\mathbf{ZAG^{-1}}$$

$$\begin{aligned}
\mathbf{F} &= \mathbf{ZAG^{-1}} \\
(\mathbf{1/N)F'F} &= (1/N)(\mathbf{ZAG^{-1}})'\mathbf{ZAG^{-1}} \\
&= (\mathbf{G^{-1}})'\mathbf{A}'(1/N)\mathbf{Z}'\mathbf{ZAG^{-1}} \\
&= (\mathbf{G^{-1}})\mathbf{A}'\mathbf{RAG^{-1}} \\
&= \mathbf{G^{-1}A'AA'AG^{-1}}\ ,\ \mathbf{A'A=G} \\
&= (\mathbf{G^{-1}G})(\mathbf{GG^{-1}}) \\
&= (\mathbf{I})(\mathbf{I}) = \mathbf{I}
\end{aligned}$$

Dabei benutzen wir A'A=G(die Summe der quadrierten Ladungen ergibt die durch die Faktoren erklärten Varianzanteile, sprich die Eigenwerte).

(18) Anteilsgrößen werden bei Korrelationsanalysen häufig transformiert, um ihre Verteilungen symmetrischer zu gestalten (siehe z. B.

Stahl 1980, S. 40). Oft wird automatisch eine Arcus-Sinus-Transformation vorgenommen. Ob diese oder eine andere Transformation überhaupt nötig oder sinnvoll ist, ist im Einzelfall oft schwierig zu entscheiden (siehe Draper/Smith 1981, S. 238 f.; Linder/Berchtold 1976, S. 19 ff.; Fox 1990, S. 105 ff.; Bacher 1990, S. 128 f.). In der folgenden Beispielanalyse wird keine Datentransformation vorgenommen. Es sei aber auf ein anderes Problem hingewiesen, das bei Anteilswerten entstehen kann, falls sie sich zu 1 (oder 100 %) ergänzen, wie z. B. die Prozentanteile der Katholiken, Protestanten und »sonstiger« Konfessionen. Diese Anteile sind linear abhängig mit der Folge: »Wenn nicht mindestens zwei dieser drei Variablen hoch korreliert sind, dann produziert diese Situation aus mathematisch-methodischen Gründen einen zweiten Faktor als Reaktion auf diesen linearen Zusammenhang, wobei dieser Faktor eigentlich nur eine kompensatorische Rolle spielt.« (Stahl 1980, S. 45)

(19) Die Berechnungen wurden mit dem ProgrammSystem SPSS^{x}, Version 4.0, durchgeführt.

(20) Die geometrische Interpretation der Faktorentransformation wird besonders sorgfältig erläutert in Bortz (1979, S. 632 ff.).

(21) Dennoch kann es sich um ein Artefakt handeln. Die Ergebnisse einer Hauptkomponenten- oder Faktoren an alyse sollten stets durch theoretische Überlegungen und eine genaue Inspektion der Korrelation smatr ix kontrolliert werden.

(22) Einzelne Variablen, deren Kommunalität erheblich unter den Kommunalitäten aller übrigen Variablen liegt, erweisen sich häufig als »Störfaktor«; in der Regel empfiehlt es sich, sie zu eliminieren (siehe Arminger 1979, S. 95)

(23) Auf den ersten Blick mögen die Faktorrotationen willkürlich und manipulativ erscheinen. Es ist jedoch daran zu erinnern, daß es gerade die Aufgabe der explorativen Faktoren an alyse ist, die Beobachtungsvariablen mit geeigneten Gewichten möglichst so zusammenzufassen, daß inhaltlich interpretierbare Faktoren »entstehen«. Die unrotierte Faktorlösung (die man benötigt, um eine Anfangsstruktur zu finden), beruht dagegen allein auf dem Kriterium maximal erklärter Varianzen. Die Rotation bzw. Transformation der Faktoren stellt nichts anderes dar als eine Umgewichtung der Ladungen der einzelnen Variablen, die linear zu einer latenten Dimension derart kombiniert werden, daß sie inhaltlich interpretier bar wird. Die Varianzsumme, die durch die transformierten Faktoren erklärt wird, bleibt unverändert.

(24) Ohne daß man die Faktorwerte errechnet hat, läßt sich die Korrelation der Faktoren mit Hilfe der Transformationsmatrix zur (schiefwinkligen) Faktorrotation berechnen.

(25) Es stellte sich heraus, daß SPSS^x, Version 4.0 diese Koeffizienten und damit die Faktorwerte falsch berechnet. Deshalb erfolgt dieser Teil der Analyse mit Hilfe des Programmpakets BMDP (in der Version von 1988).

(26) Sie wird auch als »Hauptachsenmethode« bezeichnet. Arminger (1979, S. 40) bemerkt hierzu: »Da aber die Hauptachsentransformation der linearen Algebra identisch ist mit der ... Hauptkomponentenmethode, verwenden wir den Namen Hauptfaktoren methode.«

(27) Wie oben ausgeführt, ist die Summe der Eigenwerte gleich der Summe der Kommunalitäten bzw. der erlärten Varianz.

(28) Im Falle der Hauptkomponentenanalyse führt die Regressionsschätzung zum gleichen Ergebnis wie die direkte Ableitung der Faktorwerte gemäß Gleichung (28'):

$$\begin{aligned} \mathbf{B} &= \mathbf{R}^{-1}\mathbf{A} \\ &= (\mathbf{A}\mathbf{A}')^{-1}\mathbf{A} \\ &= (\mathbf{A}')^{-1}\mathbf{A}^{-1}\mathbf{A} \\ &= (\mathbf{A}')^{-1} \end{aligned}$$

Der Schätzer B ist also identisch mit der Gewichtsmatrix der ersten Zeile in Gleichung (28').

(29) Es läßt sich zeigen, daß die iterative Hauptfaktorenmethode zu einer Lösung gemäß der Methode der Kleinsten Quadrate äquivalent ist (Arminger 1979, S. 51 f.)